An algorithm to compute data diversity index in spatial networks ☆

Taras Agryzkov[a], Leandro Tortosa[a], Jose F. Vicent[a,*]

[a]*Departamento de Ciencia de la Computación e Inteligencia Artificial, Universidad de Alicante, Campus de San Vicente, Ap. Correos 99, E–03080, Alicante, Spain*

Abstract

Diversity is an important measure that according to the context, can describe different concepts of general interest: competition, evolutionary process, immigration, emigration and production among others. It has been extensively studied in different areas, as ecology, political science, economy, sociology and others. The quality of spatial context of the city can be gauged through this measure. The spatial context with its corresponding dataset can be modelled using spatial networks. From here we can study the diversity of data present in this type of network. In this paper we propose an algorithm to measure diversity in spatial networks based on the topology and the data associated to the network. In the experiments developed with networks of different sizes, it is observed that the proposed index is independent of the size of the network, but depends on the topology of it.

Keywords: diversity index, spatial networks, urban networks, spatial statistics, Gini-Simpson index.

1. Introduction

The concept of diversity has been extensively studied in different areas. For example, in ecology, diversity characterizes communities of species and ecosystems [1, 2]. In political science, diversity is used to measure the effective number of political parties within parliament [3, 4]. Economists use diversity to measure the effective number of firms in an industry [5]. In sociology, the ethnic diversity is studied [6]. In complex network science, the diversity of node degree is used to measure an entropy of the network itself [7, 8]. In urban studies the diversity of urban facilities indicates the quality of urban environment [9, 10]. Consequently, the concept of diversity has been defined in different ways, and several different indices have been developed to express it. According to the type of sample and the application context, we can highlight the following measures of diversity: Simpson concentration index [11], Shannon index [12], Rényi entropy [13] and the set of Alfa-Beta-Gama diversities [14, 15, 16]. All of these diversities depend on the contextual characteristics of a sample, data richness and data evenness [17, 18, 19]. In this sense, the richness is a measure of the total number of data types in a sample, while the evenness expresses how evenly the individual data in a sample are distributed over the different data types.

The diversity in complex networks has been studied only from the perspective of its topology and different heterogeneity measures have been proposed in the literature to study the diversity in node degree or in the own structure of the network. But some networks, such as cities, could not be understand only from its topology. The city is a complex system [20, 21], that organizes goods and different sets of features in the service of its citizens. An appropriate analysis of these elements allows us to elaborate an idea about the quality of an urban environment. Many urban studies indicate that the quality of urban space is closely related to the diversity of services located around it. A greater mix of uses and services endows the area

☆Partially supported by the Spanish Government, Ministerio de Economía y Competividad, grant number TIN2014-53855-P.

*Corresponding author.

Email addresses: taras.agryzkov@ua.es (Taras Agryzkov), tortosa@ua.es (Leandro Tortosa), jvicent@ua.es (Jose F. Vicent)

with greater commercial activity, greater opportunities for the local business and drives a better interaction between people [9].

We have noticed that the study of data diversity in the context of the city has not been extensively studied from the perspective of spatial networks [22, 23], or more specifically, the urban networks [24], which represents the topology of the urban plots. The most common approach to the study of data diversity of the urban environment has been made using continuous samples of data. This methodology involves the division of the city area into flat regions of a certain size or grid with the parametric refinement [10]; the issue which makes the results depend on the size and relative position of the grid. These geometric dependencies, together with the extreme sensitivity of the statistical measures of diversity, can invalidate the correct interpretation of the real data diversity in the urban context. In addition, it must be taken into account that the data in the city do not appear on a continuous plane, but are constrained to a geometry of streets. Therefore, we consider that a better approach to the issue of data diversity in the urban environment would be using the spatial network model, which allows both the modelling of the topological pattern of urban streets and the georeferencing of data present in the streets of the city.

In this paper we propose a diversity index for the data associated to a spatial network. This measure is characterized by taking into account both the local data of the node and the data associated to the neighbour nodes, following paths in the network at a specified depth.

The paper is divided into the following sections. In the first section we discuss commonly used statistical measures of diversity. In the second section, we present and justify the theoretical model to measure the data diversity within the context of spatial networks. In the third section we discuss several examples where we apply the proposed model to some networks at different scales; finally, we present some conclusions.

2. Statistical models of diversity

The basic idea of a diversity index is to obtain a quantitative measure of variability that can be used to compare different samples composed of discrete components [18]. Diversity index is a characteristic value for diversity itself, which in turn can be interpreted as the effective number of types or equally abundant types in the sample [25, 26]. Diversity depends on the data richness and data evenness. Following this definition, the richness represents the number of data per sample and the evenness indicates relative abundances of the various data in a sample. The evenness increases as data are more evenly distributed in a sample such that maximum evenness is obtained when all the data are equally abundant. The value of diversity increases as the number of data per sample increases and as the abundances of data within a sample become more even [19].

Many indices for measuring data diversity have been proposed. Nearly all non-parametric diversity indices used in the sciences are monotonic functions based on the relative abundance of species $a_i = n_i/N$, with n_i referring to the abundance of the $i-th$ data in the sample, and $N = \sum_{i=1}^{R} n_i$, with R (richness) representing the total number of elements in the sample. These sets of functions include Shannon entropy, all Simpson and Gini-Simpson measures [27, 28], all Rényi entropies, all *Tsallis* entropies [29], and many others. All such measures we can obtain from the reciprocal form of the weighted power mean or Hölder mean, where the proportional abundances a_i are pounded by itself. This generalized notation for diversity has been proposed by Hill [1], and it is also known in literature as *Hill numbers* or true diversity of different orders q:

$$H_q = \left(\sum_{i=1}^{R} a_i^q \right)^{\frac{1}{1-q}}. \tag{1}$$

The parameter q controls the sensitivity of the result diversity H_q to the most and least abundant elements within the sample. Thus, with $q = 0$ the diversity H_0 is completely insensitive to the frequency of elements and comes to represent the richness R. When $q < 1$ we obtain the diversities that favour disproportionately the less frequent elements in the sample and when $q > 2$ the resulting diversities favour too much the most frequent elements of the sample [30, 31]. The most widely used diversities in many areas

of science are H_1 and H_2, that are the limit of the exponential of Shannon entropy,

$$H_1 = \exp\left(-\sum_{i=1}^{R} a_i \ln a_i\right),$$

when q approaches the unity and the reciprocal form of Simpson index or the exponential form of Rényi entropy of the second order,

$$H_2 = 1/\sum_{i=1}^{R} a_i^2,$$

for $(q = 2)$ (see [25]).

3. Diversity index in spatial networks

Our objective is to establish a model that allows us to determine an index of diversity for each of the nodes of a spatial network. As we have already mentioned, although topology is a fundamental aspect of any network, it is not enough to characterize and compare the interactions of huge and diverse data that are taking place in spatial networks.

If we analyse the heterogeneity measures proposed in literature for complex networks, it becomes clear that two different aspects of complex networks can be quantified through diversity measures. They are the diversity in node degree and the diversity in the structure of the network. What we intend in this work is to measure the diversity of nodes of the spatial network in terms of its associated data and the data topological relationships. Therefore, in order to achieve our goal we have to adapt to spatial networks the statistical models of diversity measures based on the variation of data.

In this work the nodes constitute the basic unit of information, just as the data of the node constitutes the local and exclusive sample of the node. Following this definition, node diversity can be measured at two different levels: local level and supra-local level. At the local level, the unique data of the local sample of the node is involved. At the supra-local level, both the local sample of the node itself and the local samples of adjacent nodes located at a certain topological depth are involved.

We consider that the measurement of node diversity using exclusively local data does not provide sufficient information in the context of spatial networks where the topological relationship of data constitutes the main interest of this particular type of network. In addition, applying the statistical models of diversity exclusively at the local level of the node can lead us to highly distorted diversity results regarding the real diversity of these data in the urban context. Therefore, in this work we focus on the study of diversity at an extended level which we will refer to as the *supra-local level* of the node. At this level the value of diversity depends on the local data of the node and the data of the nearest nodes.

In the study of the diversity in urban networks we consider that it is fundamental to link the network sample to some representative context of the urban environment. Therefore, we define the supra-local sample as the set of local samples of the nodes that are along the path made from the node object of the diversity measurement. Indeed, if the path is defined as the sequence of non-repetitive nodes, then the diversity of the supra-local sample of the node is given by the diversity of data that is found along the path made from that node. As we will see later, this approach allows us to weight each local sample of the path according to the topological distance between the source node and the node of the sample and also according to the degree of the nodes involved in the path.

The question that arises now is the size of the length of the paths we consider from the initial node object of the calculation. In a later section we will develop a more detailed study of this question. Now we must establish a procedure that allows us to determine the paths of length m from an initial node without repeating any node. Let us consider the problem of computing all the paths of length 2 whose origin is $v_0 = 0$ without repeating nodes. We now explain with detail this procedure:

1. We obtain a set of adjacent nodes to v_0. Let this set be:

$$\boldsymbol{v}_0^{(1)} = \{v_{01}, v_{02}, \ldots, v_{0k}\} = \{v_{0j}\}_{j=1}^{k}.$$

2. Now, we calculate the adjacent nodes of $\boldsymbol{v}_0^{(1)}$. For this purpose, for all $j = 1, 2, \ldots, k$, do the following steps:

   (a) For $j = 1$, obtain the adjacent nodes to the node v_{01}, that is,

   $$\boldsymbol{v}_{01}^{(1)} = \{v_{011}, v_{012}, \ldots, v_{01l_1}\} = \{v_{01j}\}_{j=1}^{l_1}.$$

   (b) For $j = 2$, obtain the adjacent nodes to the node v_{02}, that is,

   $$\boldsymbol{v}_{02}^{(1)} = \{v_{021}, v_{022}, \ldots, v_{02l_2}\} = \{v_{02j}\}_{j=1}^{l_2}.$$

   (c) So on until $j = k - 1$. Finally, for $j = k$, obtain the adjacent nodes to the node v_{0k}, that is,

   $$\boldsymbol{v}_{0k}^{(1)} = \{v_{0k1}, v_{0k2}, \ldots, v_{0kl_k}\} = \{v_{0kj}\}_{j=1}^{l_k}.$$

3. We construct the set of nodes adjacent to v_0 and their adjacent ones, writing them in an ordered way as

   $$\boldsymbol{v}_0^{(2)} = \{v_{01}, v_{02}, \ldots, v_{0k}, v_{011}, v_{012}, \ldots, v_{01l_1}, v_{021}, v_{022}, \ldots, v_{02l_2}, \ldots, v_{0k1}, v_{0k2}, \ldots, v_{0kl_k}\}.$$

4. Now we construct a matrix denoted by $A_{\boldsymbol{v}_0^{(2)}}$, where the columns are defined as the following criteria: the first column is the node v_0, the adjacent nodes to v_0 fulfill the following columns in an ordered form. Finally, the adjacent nodes of the adjacent ones, again in an ordered form, complete the following columns. Regarding the size of this matrix by rows, we only need to define the rows with the node v_0 (first row) and the nodes adjacent to this one, following the order previously established.

   Compute the adjacency matrix of this new matrix. In this case, it will be given by

   $$A_{\boldsymbol{v}_0^{(2)}} = \begin{array}{c} \overbrace{\begin{array}{cccccccccccccc} v_0 & v_{01} & \cdots & v_{0k} & v_{011} & \cdots & v_{01l_1} & v_{021} & \cdots & v_{02l_2} & v_{0k1} & \cdots & v_{0kl_k} \end{array}} \\ \begin{array}{c} v_0 \\ v_{01} \\ \vdots \\ v_{0k} \end{array} \left[\begin{array}{c} \\ \\ \\ \\ \end{array}\right] \end{array}. \quad (2)$$

5. Now, we determine the paths of length 2 from the matrix $A_{\boldsymbol{v}_0^{(2)}}$.

   To accomplish this task, We start by traversing the first row of the matrix $A_{\boldsymbol{v}_0^{(2)}}$. When $A_{\boldsymbol{v}_0^{(2)}}(0, i) = 1$, then we jump to traverse the i-th row of this matrix. When $A_{\boldsymbol{v}_0^{(2)}}(i, j) = 1$, we can affirm that $(0, i, j)$ constitutes a path of length 2. When we have just traversed the rows corresponding to the adjacent nodes of the origin node, we only have to make a final check. We must be sure that no path contains repeated nodes. Therefore, we check that $i \neq j$, which means that we do not pass twice by the same node. Removing the cases with repeated nodes, we already have the list with all the paths of length 2 whose origin is in v_0.

   We denote by $\mathcal{C}_{v_0}^m$ the set of all paths of length m, without repeating nodes and whose origin is the node v_0.

To better understand the model we propose to measure the nodes diversity of spatial network, we use an example of a the small graph shown in Figure 1. We represent a plane graph formed by 9 nodes with the connections given by the edges and the data associated to each one of the nodes. Each node has a sample

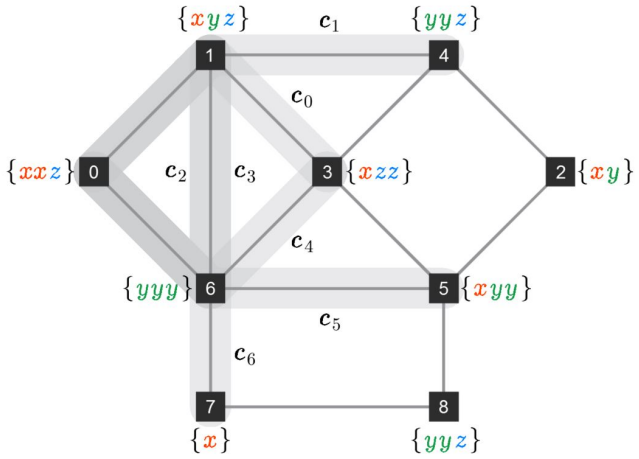

Figure 1: Example of a graph with 9 nodes, where $v_0 = 0$.

that represents these data associated to it. This sample is denoted by s, so that, for instance, the sample associated to the node v_i will be denoted by $s(i)$. The samples appear next to each node between braces, having three different types (x, y, z).

Let us take the node $0 = v_0$ as the origin node to develop the process leading us to the calculation of the diversity index. We must begin by determining the paths of length 2 with origin in v_0. Thus, we construct the following sets:

Nodes adjacent to v_0. Then, $\boldsymbol{v}_0^{(1)} = \{v_{01}(1), v_{02}(6), \}$.

Nodes adjacent to $\boldsymbol{v}_0^{(1)}$. Then $\boldsymbol{v}_0^{(2)} = \{v_{011}(3), v_{012}(4), v_{021}(7), v_{022}(5)\}$.

This way, the set of nodes adjacent to v_0 with path length equal to 2, written in an ordered form, is

$$\boldsymbol{v}_0^{(2)} = \{v_{01}, v_{02}, v_{011}, v_{012}, v_{021}, v_{022}\}\,.$$

Consequently, the matrix $A_{\boldsymbol{v}_0^{(2)}}$ is

$$A_{\boldsymbol{v}_0^{(2)}} = \begin{array}{c} \\ v_0 \\ v_{01} \\ v_{02} \end{array} \overbrace{\begin{array}{ccccccc} v_0 & v_{01} & v_{02} & v_{011} & v_{012} & v_{021} & v_{022} \\ \hline \cdot & 1 & 1 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & 1 & 1 & 1 & \cdot \\ \cdot & 1 & \cdot & 1 & \cdot & \cdot & 1 \end{array}}^{} \quad 1 \qquad (3)$$

To determine all the paths of length 2 it is enough to apply the condition

$$A_{\boldsymbol{v}_0^{(2)}}(0,i) = 1 \quad \& \quad A_{\boldsymbol{v}_0^{(2)}}(i,j) = 1 \quad \longrightarrow (0,i,j) \in \mathcal{C}^2_{v_0}.$$

Figure 2 shows us with detail, using a typical statistical tree diagram, the 7 paths of length 2 that we are able to construct whose origin is in $v_0 = 0$ and with the characteristic that no node is repeated in the path. These paths are

$$\mathcal{C}^2_{v_0} = \{c_0, c_1, c_2, c_3, c_4, c_5, c_6\}\,,$$

where each path consists of the following nodes:

$$c_0 = (0,1,3), c_1 = (0,1,4), c_2 = (0,1,6), c_3 = (0,6,1), c_4 = (0,6,3), c_5 = (0,6,5), c_6 = (0,6,7).$$

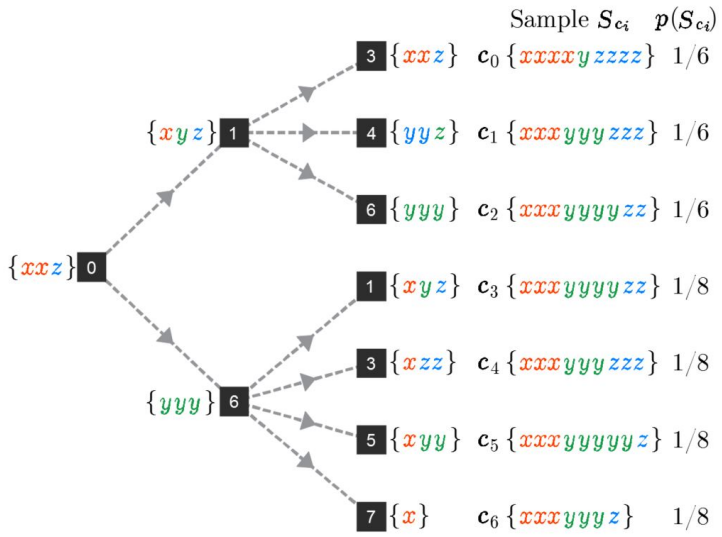

Figure 2: Source node paths from a tree diagram.

Figure 2 also shows us next to each of the nodes its local sample. In addition, the following columns in the figure show two key elements in the proposed model for calculating diversity. On the one hand, we have the total sample associated with each particular path, that is, S_{c_i}. In the following, we will refer to this total sample associated to paths as *supra-local sample*. Thus, we observe that the supra-local sample associated with the path $c_2 = (0, 1, 6)$ is formed by the sequence $\{xxxyyyyzz\}$, and it is denoted by S_{c_2}. On the other hand, we have the last column where the probability of traversing the corresponding path in the tree diagram and which constitutes the probability associated with obtaining the sample S_{c_i} appears. This is why we denote it as $p\,(S_{c_i})$.

To explain the method followed to calculate the paths from the origin node v_0 we have restricted ourselves to paths of length 2, which means that we calculate the data diversity index of v_0 taking into account the data of the nodes connected by paths of length 2 to v_0. However, we can generalize this result to paths of length m. We now consider the nodes located at a topological distance of m from v_0. Accordingly, following a scheme similar to that described above, we must calculate the sets

$$\boldsymbol{v}_0^{(1)}, \boldsymbol{v}_0^{(2)}, \boldsymbol{v}_0^{(3)}, \ldots, \boldsymbol{v}_0^{(m)}$$

With these vertices, we construct the adjacency matrix $A_{\boldsymbol{v}_0^{(m)}}$ following the same order in the placement of the elements. With the adjacency matrix we can determine all the paths of length m, which is denoted by $\mathcal{C}_{v_0}^m$. We traverse the constructed adjacency matrix and obtain the conditions:

$$\overbrace{A_{\boldsymbol{v}_0^{(m)}}(0,i) = 1 \;\&\; A_{\boldsymbol{v}_0^{(m)}}(i,j) = 1 \;\&\; A_{\boldsymbol{v}_0^{(m)}}(j,k) = 1 \& \cdots \& A_{\boldsymbol{v}_0^{(m)}}(r,s) = 1}^{m \quad \text{conditions}}$$

Therefore, after checking that we have no path repeating two nodes (if there are any they must be removed), we have that

$$(0, i, j, k, \ldots, r, s) \in \mathcal{C}_{v_0}^m.$$

After completing the first phase of the model consisting of the determination of all the paths of length m whose origin is v_0, we can already consider how we can calculate the index of diversity associated with the node. The main idea is to apply a diversity index to each of the supra-local samples obtained S_{c_i}, which will be multiplied by a factor (more precisely, by the probability $p(S_{c_i})$). Once this calculation has been

carried out, it is enough to add the values of all the supra-local samples to obtain the diversity index for the node v_0. We analyse this process in detail.

The main advantage of this approach is that it allows us to combine statistical models of data diversity with the probabilistic model of supra-local samples of the nodes proposed in this work.

Among all the diversity measures derived from Hill numbers (1), we are especially interested in Gini-Simpson index or also known as *Probability of Interspecific Encounter* (PIE) [32], which can be expressed for continuous data in a sample by using the true diversity expression (1), for $q = 2$, as

$$I = 1 - \frac{1}{H_2} = 1 - \sum_{i=1}^{R} a_i = 1 - \sum_{i=1}^{R} \left(\frac{n_i}{N}\right)^2.$$

Therefore, we arrive to the expression

$$I_{GS} = 1 - \sum_{i=1}^{R} \left(\frac{n_i}{N}\right)^2. \tag{4}$$

for the Gini-Simpson diversity index.

The original Simpson index (measure for concentration of dataset) represents the probability that two entities selected at random from the dataset without replacement represent the same type [19, 31]. Its mathematical expression is

$$I_S = \sum_{i=1}^{R} \frac{n_i\,(n_i - 1)}{N\,(N - 1)}. \tag{5}$$

Following this definition, the less diverse the dataset is, the higher is the value I_S. Its transformation, commonly known as Gini-Simpson Index $1 - I_S$, represents the probability that two entities belong to different types. It should be noted that for continuous samples this model is usually used, but with data replacement, as it is expressed in (4). The most notable difference between the two expressions is that expression (5) does not depend on the sample richness R, its maximum value will be equal to 1 as long as all species are equally abundant. Nevertheless, the value obtained by the expression (4) depends on both, the richness of the sample and the evenness of types, and generally works better for large samples [18], as it is in our case with the introduction of the supra-local samples. This leads us to choose the Gini-Simpson diversity index for this model.

Consequently, the diversity index for spatial networks proposed in this paper combines the statistical model of Gini-Simpson diversity index (4) with the probabilistic process of calculation of the diverse paths that depends exclusively on the topology of the network.

The process to compute the diversity index of a node v_0 using paths of length m, denoted by $D_{v_0}^m$ in a spatial network may be summarised in the three following steps:

1. For each supra-local sample obtained along the probable path of length m from the node origin v_0, we calculate the probability that two randomly chosen elements from the supra-local sample (with replacement) are of different types.

2. The obtained result of the first step is weighted with the probability of the path itself.

3. The diversity of the node is given by the addition of the results obtained for each path.

This process can be mathematically formulated by the expression

$$D_{v_0}^m = \sum_{c_i \in \mathcal{C}_{v_0}^m} p\left(S_{c_i}\right) \cdot I_{GS}\left(S_{c_i}\right), \tag{6}$$

where $p\left(S_{c_i}\right)$ represents the probability associated to the supra-local sample S_{c_i} and $I_{GS}\left(S_{c_i}\right)$ represents the Gini-Simpson diversity index of the supra-local sample S_{c_i} (associated to the path c_i). In (6) we are assuming that the length of the paths from v_0 is m.

For cases where nodes are empty of data, we have that the diversity index is null.

It should be noted that the idea of introducing the multiplicative factor of probability to the sample of each path is fundamental in our model and supposes a novel analysis in this problem. We might think, when considering paths of a certain length from the origin node v_0, that the sample associated with the starting node is equally important than the rest of the samples involved in the general calculation. However, this is not the case because when introducing the probability associated with each path, some samples appear many times (much more than others) when adding the samples. More specifically, the samples closest to the origin node have a much larger weight in the overall calculation than those that are more distant from the origin.

Path c_i / Sample s_i	s_0	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8
c_0	1/6	1/6	0	1/6	0	0	0	0	0
c_1	1/6	1/6	0	0	1/6	0	0	0	0
c_2	1/6	1/6	0	0	0	0	1/6	0	0
c_3	1/8	1/8	0	0	0	0	1/8	0	0
c_4	1/8	0	0	1/8	0	0	1/8	0	0
c_5	1/8	0	0	0	0	1/8	1/8	0	0
c_6	1/8	0	0	0	0	0	1/8	1/8	0
Sample s_i weight	1.000	0.625	0.000	0.292	0.167	0.125	0.667	0.125	0.000

Table 1: Paths and samples for the example studied.

To better understand this effect, we have constructed Table 1 which shows the weight of each of the samples in each of the paths, referring to the example studied above. Thus, for example, we clearly observe the weight of the sample associated with the 0 node, that is, s_0, which appears on all paths and whose total weight is 1. The sample s_1 associated with the node 1 only appears in four paths (c_0, c_1, c_2, c_3), so its weight drops to 0.625 from the sample s_0. We note that the sample of node 2 has a weight of 0. This is absolutely expected since the node 2 does not appear in any path of length 2 from the origin 0, as well as it happens with the node 8. This example clearly demonstrates the idea of giving greater importance in the calculation of diversity to the samples closest to the origin node. In addition, with this we are also attaching great importance to the network topology, since the degrees of the nodes are an essential factor in the whole process.

4. An algorithm to compute the diversity in spatial networks

From the description of the model that has been made in the previous section, we can establish the following algorithm, which summarizes the steps necessary to calculate the diversity index of any node in a spatial network.

Let us suppose that we want to determine the diversity index of any node v_0 of a spatial network composed of N nodes v_i, for $i = 0, 1, 2,\ ldots, N-1$. Let us also suppose that we are going to extend the influence of the data of the node v_0 to the nodes that are at a topological distance of m following paths of length m without repeating nodes. The spatial network has associated a set of data that have been assigned to the basic units of information that are the nodes. Thus, each node in the network has a data sample where the type of data and its quantity are shown. This sample associated with the node v_i is denoted as s_i and it is interpreted as a set of local data of the node.

Taking all these considerations, the following algorithm provides the diversity index $D^m_{v_0}$ of the node v_0.

Algorithm 1. ***Algorithm to determine*** $D^m_{v_0}$.

Let us assume that we have a graph $G = (V, E)$ representing an urban network with N nodes v_i. Let us assume that every node v_i has a dataset or sample associated to it that we will denote as s_i. We proceed with the following steps.

Step 1 *Obtain the adjacency matrix A from the primal graph of the network.*

Step 2 *Taking v_i as the origin, determine the set of adjacency nodes*

$$\boldsymbol{v}_i^{(1)}, \boldsymbol{v}_i^{(2)}, \boldsymbol{v}_i^{(3)}, \ldots, \boldsymbol{v}_i^{(m)}.$$

Step 3 *Construct the adjacency matrix $A_{\boldsymbol{v}_i^{(m)}}$ constructed from the adjacency nodes.*

Step 4 *Calculate all the paths of length m, that is, $\mathcal{C}_{v_i}^m$.*

Step 5 *For every $c_i \in \mathcal{C}_{v_i}^m$, we calculate*

(a) $p(S_{c_i})$, the probability of walking the path c_i within the network.

(b) S_{c_i}, the sample associated to the path c_i.

Step 6 *Calculate $I_{GS}(S_i)$, the Gini-Simpson diversity index for the sample S_{c_i} according to the expression (4).*

Step 7 *Calculate the diversity index $D_{v_i}^m$ for the node v_i by the expression (6).*

The algorithm summarizes the steps developed in the model, based on the calculation of the diversity of a sample, the Gini-Simpson index, although the model presented through this algorithm surpasses the idea of local diversity to take into account the relationships between the elements of a network, especially those that are connected.

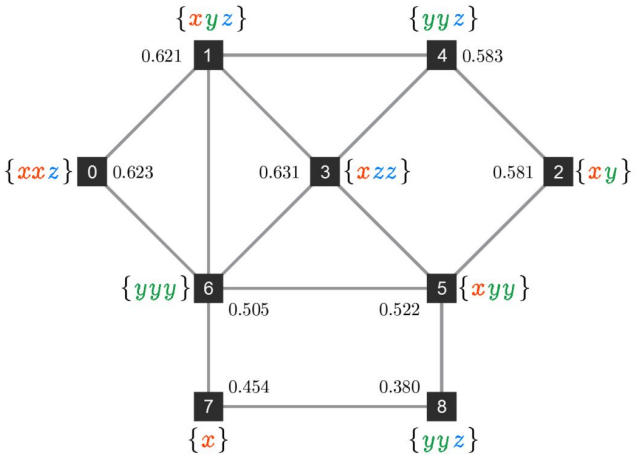

Figure 3: Small example with the diversity index for each node.

Following the example we have been studying, we have calculated the diversity indexes of each of the nodes for paths of length 2, as can be seen in Figure 3.

5. An example of a small urban network analysing the depth effect

Let's start with a very simple example in which we try to show that using data exclusively from the local samples of the nodes can lead us to obtain some results of diversity index very far from the real data diversity.

Therefore, consider a small network with 4 nodes represented in Figure 4(a), where each of the nodes have associated some data of the same type. Then, when calculating the diversity of each node using exclusively

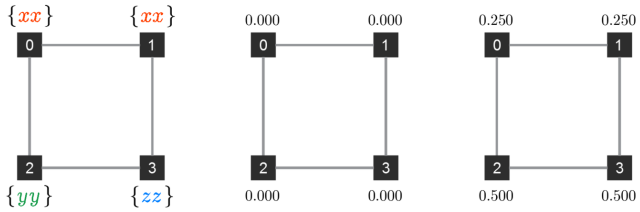

Figure 4: Simple graph with four nodes.

its local data, we obtain that the diversity index is 0 for each one of the nodes, as we can see in Figure 4(b). This does not seem consistent if the physical context is considered, especially if the nodes are close to each other.

The results obtained after applying the algorithm proposed in this paper are quite different, as we can check in Figure 4(c), where the diversity index becomes more consistent according to the data allocated in the network.

In the following, we consider an example of a fictitious 20-node urban network shown in Figure 5. The model used for the representation of the urban network is the well-known *primal graph*. The fundamental topological characteristic of this type of graph is its homogeneous degree distribution, which usually does not exceed the value of 6. We have assigned data to each node of the network in a random way, simulating the common data present in the city. As we can see from Figure 5, the associated dataset has the following types: x, y and z with the richness $0 \leq R \leq 3$ of each local sample.

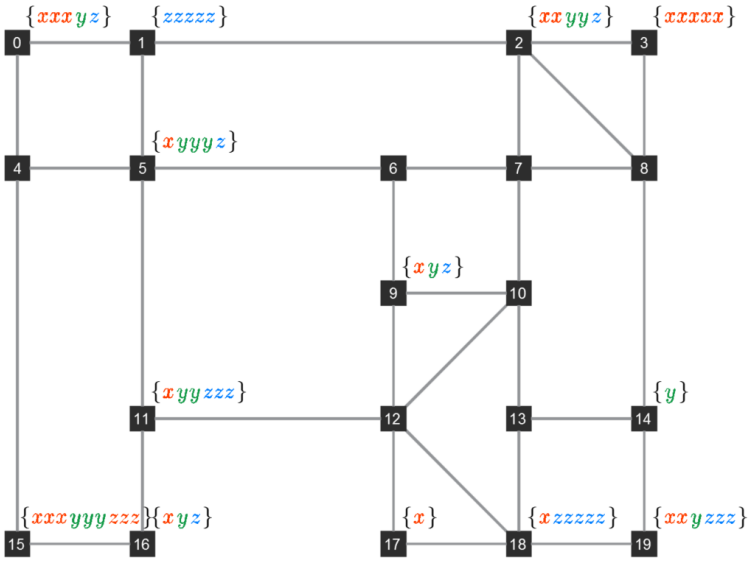

Figure 5: An example of a 20-node urban network.

We apply Algorithm 1 to the graph in Figure 5, considering the paths of all possible lengths, from $m = 0$ to $m = 19$. Notice that for $m = 0$ we have the particular case in which we apply the algorithm to the individual data of each node. In this case, the obtained result (see Figure 6a) is equivalent to the direct application of the Gini-Simpson formula 4 to each node of the network. In order to represent and visualise the obtained numerical values of diversity D we use the HotCold Scale, based on the RGB model.

The rest of the images appearing in Figure 6 correspond to different applications of Algorithm 1 to the

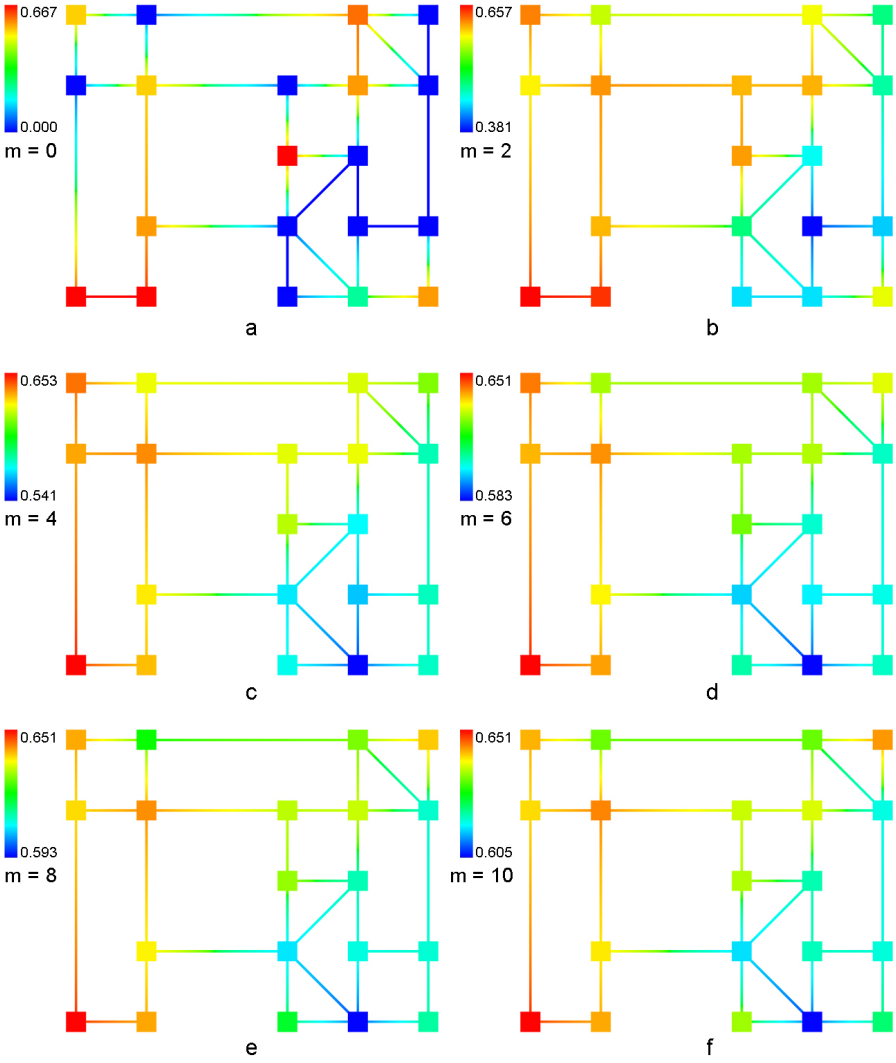

Figure 6: Numerical results of D for the studied network using different m values.

studied network, varying the values of m, that is, increasing the size of the supra-local sample S.

In the light of the results it is observed that for $m = 0$ the values of the diversity D for adjacent nodes present a great variation, which is obvious, because the topology of the network is not taken into account. Now, as we progressively increase the value of m, this effect is smoothed, until arriving to $m = 6$ where we can already find areas in the network with a more homogeneous diversity, as it is shown in Figure 6e.

The question that arises now is what length of the path m should be taken in the study of the data diversity D in spatial networks. So, in the following, taking the above example, we analyse how the values of the node diversity D are affected when varying m. For this purpose, the graphic in Figure 7 has been elaborated. In this graphic the horizontal axis represents paths of length between 0 and 19, while the vertical axis represents the value of the diversity for each of the nodes. We can interpret this graphic as a matrix where the element (i, j) represents the diversity value D of node i for a path of length j.

As in the previous graphic of Figure 6 we use the similar chromatic scale in order to visualize the variation of diversity D for different values of m. This graphic very clearly shows the evolution of the node diversity values as the size of the supra-local sample or the path length increases. So, analysing these variations, we

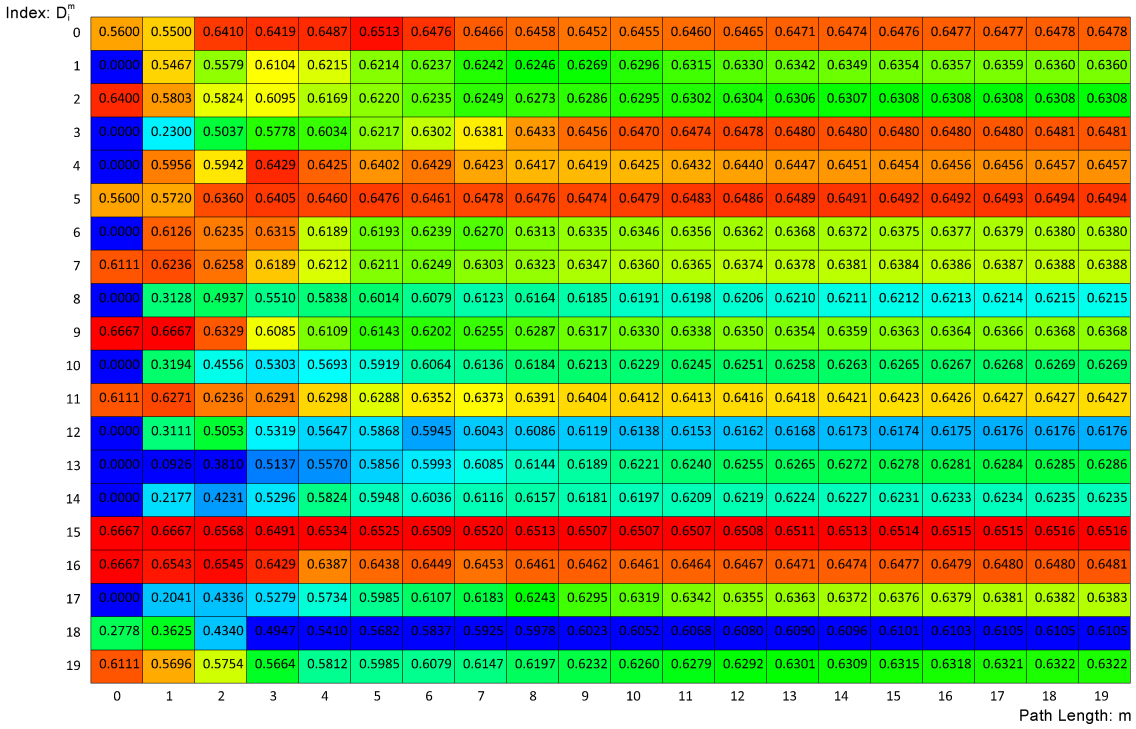

Figure 7: Node diversity D for each path length m.

conclude that for $m \geq 6$ there are no significant changes in the nodes diversity. To explore this issue further, we propose the graphic of Figure 8, where we represent in the horizontal axis the length of the path m and in the vertical axis we represent the value

$$\| \boldsymbol{x}_i - \boldsymbol{x}_{i-1} \|,$$

where $\boldsymbol{x}_i$ is the vector that contains the diversity values for the path length i of all the nodes of the network.

Thus, we see clearly how the calculated differences decrease as the path length increases. The curve of the graphic converges and from $m = 6$ it approaches notably to 0; this fact reinforces the idea expressed previously in the graphic of Figure 8.

We have performed numerical tests taking other graphs of the same type with a significantly higher number of nodes and we have obtained similar results. Therefore, we can say that the most adequate length of m is independent of the size of the network, but rather depends on its topology, i.e., on the node degree distribution. Thus, for the primal graphs representing the urban networks it is convenient that $m \leq 6$.

6. Conclusions

In this paper we propose the diversity index applicable to the data present in spatial networks. This index presents the characteristic of combining the statistical model based on the Gini-Simpson formula with a topological model based on considering the paths of variable length. Thus, the proposed diversity index of a node takes into account both the local sample data and the sample data of its neighbour nodes. An algorithm that calculates the diversity of data for each node using a function of path length has been implemented for spatial networks.

Numerical experiments have been developed to study how the size of the supra-local sample (the path length) affects the value of diversity. The study was carried out taking as an example an urban network because it is a characteristic type of spatial network in which we can associate directly the data present in

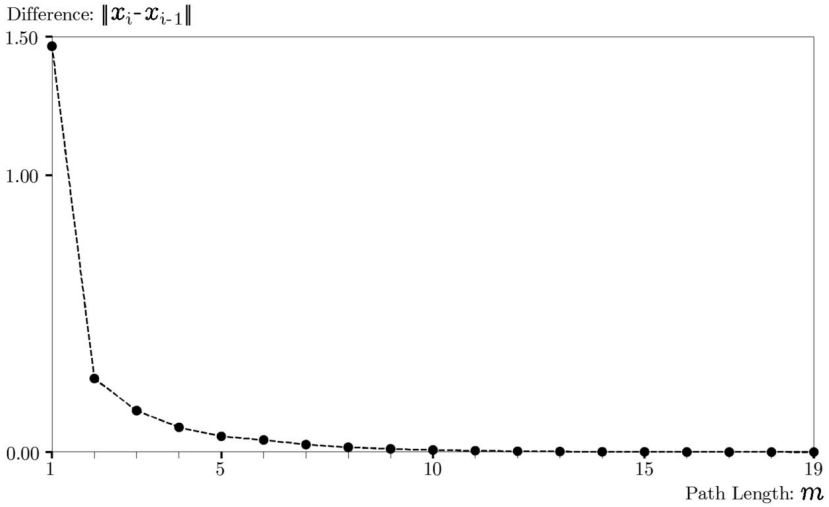

Figure 8: Differences against to the path length.

spatial context; this is the case of the urban environment. These numerical tests lead us to determine a range of values suitable for the size of supra-local sample within this type of network. In the experiments carried out with other networks of different sizes, it is observed that the proposed index is independent of the size of the network, but depends on the topology of the network, which is an important aspect from a computational point of view. The proposed measure is also applicable to other types of spatial networks, as long as they have the capacity to be carriers of information.

References

[1] Mark O Hill. Diversity and evenness: a unifying notation and its consequences. *Ecology*, 54(2):427–432, 1973.

[2] Carlo Ricotta. Parametric scaling from species relative abundances to absolute abundances in the computation of biological diversity: a first proposal using shannons entropy. *Acta Biotheoretica*, 51(3):181–188, 2003.

[3] Markku Laakso and Rein Taagepera. effective number of parties: a measure with application to west europe. *Comparative political studies*, 12(1):3–27, 1979.

[4] Jean-François Caulier and Patrick Dumont. The effective number of relevant parties: how voting power improves laakso-taageperas index. (17846), 2005.

[5] Leslie Hannah and John Anderson Kay. *Concentration in modern industry: Theory, measurement and the UK experience*. Springer, 1977.

[6] Jaana Juvonen, Adrienne Nishina, and Sandra Graham. Ethnic diversity and perceptions of safety in urban middle schools. *Psychological Science*, 17(5):393–400, 2006.

[7] Ernesto Estrada. Generalization of topological indices. *Chemical physics letters*, 336(3):248–252, 2001.

[8] Ernesto Estrada. *The structure of complex networks: theory and applications*. Oxford University Press, 2012.

[9] Salvador Rueda. La ciudad compacta y diversa frente a la conurbación difusa. *Ciudades para un futuro más sostenible*, pages 69–80, 1997.

[10] Salvador Rueda. Modelos e indicadores para ciudades más sostenibles. In *Economía, ecología y sostenibilidad en la sociedad actual*, pages 115–154. Fundación Universidad de Verano de Castilla y León, 2000.

[11] Edward H Simpson. Measurement of diversity. *Nature*, (163), 1949.

[12] Claude Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423 and 623–656, 1948.

[13] Alfréd Rényi. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 547–561, 1961.

[14] J John Sepkoski. Alpha, beta, or gamma: where does all the diversity go? *Paleobiology*, 14(03):221–234, 1988.

[15] Hanna Tuomisto. A diversity of beta diversities: straightening up a concept gone awry. part 1. defining beta diversity as a function of alpha and gamma diversity. *Ecography*, 33(1):2–22, 2010.

[16] Hanna Tuomisto. A diversity of beta diversities: straightening up a concept gone awry. part 2. quantifying beta diversity and related phenomena. *Ecography*, 33(1):23–45, 2010.

[17] Robert K Peet. The measurement of species diversity. *Annual review of ecology and systematics*, 5:285–307, 1974.

[18] Carlo Heip, Peter Herman, and Karline Soetaert. Indices of diversity and evenness. 24(4):61–87, 1998.
[19] Theodore M. DeJong. A comparison of three diversity indices based on their components of richness and evenness. *Oikos*, 26(2):222–227, 1975.
[20] Michael Batty. *Cities and complexity: understanding cities with cellular automata, agent-based models, and fractals*. The MIT press, 2007.
[21] Juval Portugali. *Complexity, cognition and the city*. Springer Science & Business Media, 2011.
[22] Dimitri Volchenkov and Philippe Blanchard. Random walks along the streets and canals in compact cities: Spectral analysis, dynamical modularity, information, and statistical mechanics. 75(2):026104, 2007.
[23] Marc Barthlemy. Spatial networks. 499(1):1–101, 2011.
[24] Bin Jiang. A topological pattern of urban street networks: universality and peculiarity. 384(2):647–655, 2007.
[25] Lou Jost. Entropy and diversity. *Oikos*, 113(2):363–375, 2006.
[26] Sönke Hoffmann and Andreas Hoffmann. True diversities: A comment on lou josts entropy and diversity. *Working Paper*, 2006.
[27] Covadonga Caso and Maria Angeles Gil. The gini-simpson index of diversity: estimation in the stratified sampling. *Communications in Statistics-Theory and Methods*, 17(9):2981–2995, 1988.
[28] Radu Cornel Guiasu and Silviu Guiasu. The weighted gini-simpson index: revitalizing an old index of biodiversity. *International Journal of Ecology*, 2012:10 pages, 2012.
[29] Constantino Tsallis. Possible generalization of boltzmann-gibbs statistics. *Journal of statistical physics*, 52(1):479–487, 1988.
[30] Constantino Tsallis. Nonextensive statistical mechanics and thermodynamics: Historical background and present status. In *Nonextensive statistical mechanics and its applications*, pages 3–98. Springer, 2001.
[31] Chris Keylock. Simpson diversity and the shannonwiener index as special cases of a generalized entropy. *Oikos*, 109(1):203–207, 2005.
[32] Stuart H Hurlbert. The nonconcept of species diversity: a critique and alternative parameters. *Ecology*, 52(4):577–586, 1971.